DOCUMENT RESUME

ED 452 211 TM 032 496

AUTHOR Hess, Brian; Olejnik, Stephen; Huberty, Carl J.

TITLE The Efficacy of Two Improvement-over-Chance Effect Size Measures for Two-Group Univariate Comparisons under Variance Heterogeneity and Nonnormality.

PUB DATE 2001-04-10

NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Comparative Analysis; *Effect Size; *Regression (Statistics); *Sample Size

IDENTIFIERS *Logistic Regression; *Predictive Discriminant Analysis

ABSTRACT

The efficacy of two improvement-over-chance or "I" effect sizes, derived from predictive discriminant analysis (PDA) and logistic regression analysis (LRA), were investigated for two-group univariate mean comparisons. Data were generated under selected levels of population separation, variance pattern, sample size, and distribution shape. Based on the accuracy of sample estimates, both "I" indices are acceptable under optimal conditions except when both population separation and sample size are small. Under variance heterogeneity and normality, "I" derived from LRA is acceptable if "n" sizes are equal. When "n" sizes are unequal, "I" derived from LRA is acceptable only if variance heterogeneity is moderate and population separation is not small. Under nonnormality, "I" derived from LRA is acceptable regardless of the variance pattern provided "n" sizes are equal. Finally, for greater precision, "I" derived from LRA should be used under large sample sizes. Some practical implications are provided. (Contains 5 tables, 4 figures, and 43 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running Head: IMPROVEMENT-OVER-CHANCE EFFECT SIZE

ED 452 211

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B. Hess

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

# The Efficacy of Two Improvement-Over-Chance Effect Size Measures for Two-Group Univariate Comparisons under Variance Heterogeneity and Nonnormality

Brian Hess, Stephen Olejnik, and Carl J Huberty
The University of Georgia

TM032496

Paper Presented at the Annual Meeting of the American Educational Research Association, Seattle, WA. April 10, 2001.

Correspondence concerning this paper may be sent to Brian Hess, 12711 Arbor Isle Drive, Temple Terrace, FL 33637. Email: bhess@coe.uga.edu.

Abstract

The efficacy of two improvement-over-chance or I effect sizes, derived from predictive discriminant analysis (PDA) and logistic regression analysis (LRA), were investigated for two-group univariate mean comparisons. Data were generated under selected levels of population separation, variance pattern, sample size, and distribution shape. Based on the accuracy of sample estimates, both I indices are acceptable under optimal conditions except when both population separation and sample size are small. Under variance heterogeneity and normality, I derived from LRA is acceptable if n sizes are equal. When n sizes are unequal, I derived from LRA is acceptable only if variance heterogeneity is moderate and population separation is not small. Under nonnormality, I derived from LRA is acceptable regardless of the variance pattern provided n sizes are equal. Finally, for greater precision, I derived from LRA should be used under large sample sizes. Some practical implications are provided.

# Efficacy of Two Improvement-Over-Chance Effect Sizes for Two-Group Univariate Comparisons under Variance Heterogeneity and Nonnormality

Hypothesis testing based on statistical inference has been the dominant data analysis method for the development of knowledge in the social sciences. Despite its dominance, many have criticized its use (e.g., Carver, 1978, 1993; Falk & Greenbaum, 1995; Huberty & Pike, 1999; Schmidt, 1996). It is recognized that hypothesis testing has limitations. For example, statistically significant p values do not imply meaningfulness and therefore do not sufficiently describe mean comparison assessments. Emphasizing p values alone may lead to poor decision making in the form of reporting trivial effects due to large sample sizes. Consequently, the move toward measuring and reporting effect sizes has gained attention and momentum (e.g., see Greenwald, Gonzalez, Harris, & Guthrie, 1996; Kirk, 1996; Olejnik & Algina, 2000; Richardson, 1996; Strube, 1988; Thompson, 1999a, 1999b). In fact, a report from Wilkinson and the APA Task Force on Statistical Inference (1999) recommended that researchers "always report effect size measures for primary outcomes (p. 599)."

Two popular approaches for estimating the magnitude of an effect are the standardized mean difference, $\delta$, (Cohen, 1988; Glass, 1977; Hedges, 1981) and measures of association such as $\eta^2$ and $\omega^2$ (Olejnik & Algina, 2000; Richardson, 1996). One common feature of these effect size measures is that they assume homogeneity of population variances. Wilcox (1987) noted that if this assumption is violated the standardized mean difference provides no pure measure of effect. Carroll and Nordholm (1975) showed the limitations of measures of association when sample sizes are unequal and variances are heterogeneous. Therefore, what is needed is an index

that can quantify practical significance in such a way that can be interpreted meaningfully when population variances differ.

Distribution Overlap

Huberty and Lowman (2000) proposed an index for quantifying practical significance under variance heterogeneity. Their effect size is obtained by way of a classification analysis (i.e., predictive discriminant analysis) and is based on the overlap of the continuous outcome variable score distributions for the groups under study. The concept of group overlap in the behavioral sciences dates back to Tilton (1937) and was revisited by Alf and Abrahams (1968), Elster and Dunnette (1971), Huberty and Holmes (1983), and Levy (1967) relating the concept of group overlap to two-group mean differences testing.

The effect size measure proffered by Huberty and Lowman (2000) was developed from an earlier investigation by Huberty and Holmes (1983) who discussed the use of univariate classification as a way to assess two-group comparisons, and percent group distribution overlap was thus presented in terms of classification proportions. Assuming that the two score distributions are similar and normally distributed, the two group means are different if the group overlap is small. One reasonable assessment approach to determine percent of group overlap is to use a univariate group membership prediction (or classification) rule. Two approaches that might be used for developing classification rules are (1) predictive discriminant analysis (PDA), and (2) logistic regression analysis (LRA) for the two-group comparison context.

Indexing Percent of Distribution Overlap using PDA

Hit rates. The amount of group overlap is determined by calculating an across-group membership hit rate. A hit rate is the proportion of analysis units that are correctly assigned to the group from which they emanate. The assignment to groups is based on a classification rule.

In the univariate case, if population variances can be assumed to be equal, then a linear rule may be used. Under this rule sample variances are pooled when computing posterior probability estimates of group membership, $P(g \mid X_i)$. These posterior probability estimates reflect the probability that the ith unit will belong to population g, given an observed score $X_i$. The linear classification rule can be obtained by

$$P(g \mid X_i) = \frac{q_g \cdot \exp(-1/2\ D^2_{ig})}{\sum^{k}_{g'=1} q_{g'} \cdot \exp(-1/2\ D^2_{ig'})} \quad . \tag{1}$$

In Equation 1, $D^2_{ig}$ is the Mahalanobis squared distance of unit i from the mean of group g, [or, $(X_i - X_g)'s^{-1}(X_i - X_g)$, where $X_g$ is the mean of group g and $s^{-1}$ is the pooled variance on the predictor variable], and $q_g$ is the probability that any unit is a member of population g. The prior probabilities reflect the relative sizes of the populations involved in the group comparisons. For example, in an experimental context where individuals are randomly assigned to a treatment or control condition, it is reasonable to set $q_1$ and $q_2$ equal to .5 because the probability of a unit belonging to one of the two groups is equally likely. Conversely, in a nonexperimental study where random assignment to groups is not possible (e.g., ethnicity) it is important to choose probabilities that reflect the population proportions in order to obtain an appropriate across-group hit rate.

If population variances cannot be assumed to be equal, then a quadratic rule would be used. The quadratic classification rule can be obtained by

$$P(g \mid X_i) = \frac{q_g \cdot s_g^{-1/2} \cdot \exp(-1/2\ D^2_{ig})}{\sum^{k}_{g'=1} q_{g'} \cdot s_{g'}^{-1/2} \cdot \exp(-1/2\ D^2_{ig'})} \quad . \tag{2}$$

Unlike the linear rule, the quadratic rule shown in Equation 2 uses separate variances, $s_g$. In practice, equality of population variances (and covariances) can be assessed statistically by a $\chi^2$

(Bartlett) statistic or by an approximate F (Box) test. However, these tests are sensitive to distributional nonnormality (Huberty, 1994, pp. 63-64). That is, the null hypothesis of equal population variances/covariances might be rejected due to nonnormality and not because of true variance/covariances inequality.

Huberty (1994, p. 87) recommended an external classification analysis to determine an across-group classification or hit rate. In an external analysis, the classification rule is determined on one set of units and then used to classify another set of units (Huberty, 1994, p. 87). One way of carrying out an external analysis involves sample splitting. One hit rate estimation technique carried out this way is termed leave-one-out (L-O-O) (see Huberty, 1994, pp. 89-93). According to Huberty and Lowman (2000), an across-group hit rate estimate using the L-O-O approach and counting the number of units correctly classified yields a good representation of group overlap. That is, the L-O-O method will yield an acceptable point estimate of the true across-group hit rate.

Improvement-over-chance. After the across-group hit rate estimate is calculated, an estimate of the magnitude of the effect can be obtained. Huberty and Lowman (2000) pointed out that an estimated across-group observed hit rate, denoted as $H_o$, by itself might not be an adequate effect size index. If the observed hit rate is high but only slightly better than what one may expect by chance, then the effect would not be very impressive. Under a proportional chance criterion, the expected or chance frequency of correct classification for group g is $e_g = q_g n_g$, where $q_g$ is the prior probability for group g, and $n_g$ is the number of analysis units in group g. The expected or chance frequency of correct classification across groups then is $e = \Sigma e_g$. From this the expected or chance hit rate across groups is $H_e = e / N$.

Huberty (1994, p. 107) proposed a reduction-in-error or improvement-over-chance ( I ) index:

$$I = \frac{H_o - H_e}{1 - H_e} \quad . \tag{3}$$

I indicates the proportion of correct classification (or hit rate) that is less than that made if classification were done by chance. In other words, Equation 3 addresses the question, "To what extent is the group distribution overlap greater than what may have been expected by chance alone (i.e., due random classification)?"

Huberty and Lowman (2000) provided some preliminary evidence as to the efficacy of I as a measure of effect size for univariate and multivariate group comparisons. Using extant data, they compared I to the point-biserial correlation ($r_{pb}$), F, and $\eta^2$. In the two-group comparison case, they found that the relationship between $r_{pb}$ and I was .90. For the $k>2$-group homogeneity of variance condition, the relationship between F and I was .93 and between $\eta^2$ and I was .97. When the variances were not judged to be homogenous, I was compared to adjusted F values (or J values) based on the James Second order test (Oshima & Algina, 1992). Using a quadratic rule to obtain I values, they found that the correlation between J and I values was .89. Based on these preliminary analyses, Huberty and Lowman concluded that I may be used in situations that are univariate, multivariate, homogeneous, heterogeneous, or any combination thereof.

Logistic Regression Analysis

Another popular method for two-group classification is logistic regression analysis (LRA). Whereas discriminant analysis is part of the general linear model, logistic regression models the nonlinear probabilistic function of the dichotomous variable (Fan & Wang, 1999).

Computationally, obtaining hit rates using LRA is intuitively simpler than PDA. Given a binary (dichotomous) outcome variable Y (Y = 0 or 1), such as group membership in the two-

group context, and a single predictor (continuous) variable X, the posterior probability of belonging to the target group (e.g., $\underline{Y} = 1$ for group 1 membership) is modeled through the logistic function

$$\underline{Y} = \frac{e^{(\beta' X)}}{1 + e^{(\beta' X)}} . \quad (4)$$

Assuming only one predictor in Equation 4, $\beta' X = \beta_0 + \beta_1 X_1$ and $\underline{Y}$ is the predicted posterior probability of an observation belonging to group 1. Unlike parameter estimates in PDA, estimates of logistic regression model parameters ($\beta'$) cannot be obtained analytically. Consequently, maximum likelihood estimators for $\beta'$ are obtained iteratively.

Once the logistic regression model is established, the model may be used to obtain the classification or hit rate. In doing so, obtaining an observed hit rate is straightforward: Classify $X_i$ into the target group (group 1) if the predicted posterior probability of the observation for that group is large, otherwise classify the observation into the other group. The problem, however, is to determine the cutoff point for the predicted probability above which $X_i$ will be classified into the target group, and below which $X_i$ will be classified into the other group. Typically, the specific cutoff value is based on the size of the modeled population.

Like PDA, it seems reasonable that the amount of overlap between two population distributions can be assessed similarly using LRA, and subsequently, values of $\underline{I}$ can be obtained using the estimated hit rates computed from LRA. The question then is, "Under which conditions of variance heterogeneity might one use LRA over quadratic PDA as the method to compute $\underline{I}$?"

PDA vs LRA. Both PDA and LRA can be used to compute $\underline{I}$. Theoretically, in the two-group context, hit rates obtained from linear PDA are identical to the hit rates obtained from

LRA regardless of the variance pattern (Efron, 1975). Because LRA is relatively free of stringent data conditions it is viewed as a more flexible method (Cox & Snell, 1989; Fan & Wang, 1999; Neter, Wasserman, & Kutner, 1989).

The relative performance of PDA compared to LRA for two-group classification has been extensively studied (e.g., Dattalo, 1994; Efron, 1975; Fan & Wang, 1999; Meshbane & Morris, 1996; Press & Wilson, 1978). When PDA assumptions are met very little difference in classification accuracy have been observed (Fan and Wang, 1999). Similarly, when the two groups had unequal covariance matrices and very different n sizes, the classification rates for both PDA and LRA for the total sample (across-groups) were comparable. In addition, Fan and Wang found that sample size played a minor role in the classification accuracy of the two methods; however LRA did require larger sample sizes to achieve stable classification results.

When Fan and Wang (1999) computed hit rates using PDA under variance/covariance heterogeneity, they used a linear rule (pooled the covariance matrices) to obtain the linear classification function values. These linear PDA hit rates were compared to those based from LRA. Subsequently, the comparability of quadratic PDA and LRA under variance heterogeneity was not studied. As previously noted, when population variance/covariance matrices are judged to be heterogeneous, using a linear rule in PDA is not appropriate, and therefore a quadratic rule should be used (Huberty & Lowman, 2000). Thus, there is no evidence to date that would suggest the superiority of either quadratic PDA or LRA as methods for computing I under variance heterogeneity.

Given the limited understanding and application of the I index as a measure of effect size used in mean comparison assessments, little is known about its sampling properties under various data conditions. In the context of two-group univariate mean comparisons, the

distributional properties of I using linear and quadratic PDA or LRA have not been studied. In addition, the comparability of the two methods on which I is based is unknown. Therefore, the purposes of the present study were to (1) describe the sampling characteristics of I derived from both linear and quadratic PDA and LRA as a measure of effect size for two-group univariate mean comparisons under relevant data conditions, and (2) provide recommendations for using either PDA or LRA for deriving I, particularly if quadratic PDA or LRA should be used under variance heterogeneity (and nonnormality).

## Method

The following four data conditions were manipulated to study the sampling characteristics of each I index: (1) population separation (effect size), (2) variance pattern, (3) total sample size with equal and unequal n, and (4) distribution shape. Variations in these data conditions are commonly found in social science literature and in most practical situations; previous simulation studies have found these to be critical determinants of understanding the sampling properties of the F and t statistics (Harwell, Rubenstein, Hays, & Olds, 1992).

Three levels of population separation, or $\delta$, were considered. These $\delta$ values were set so that population 2 had a mean which was .2, .5, or .8 standard deviations greater then population 1 ($\sigma = 1$). These were chosen based on relative values of d outlined by Cohen (1988, pp. 24-27) as "small," "medium," and "large" benchmark effect sizes; these benchmarks are also embraced by some social scientists in practice. Fowler (1988) considered somewhat similar levels but extended the number of levels to include 1.0 and 1.5.

Three population variance ratios were considered: 1:1, 1:4, and 1:8. These variance patterns reflect a consistent variance of 1 for population 1 while the variance of population 2 is incremented 1, 4, and 8. Previous researchers have used similar variance patterns and the 1:4

ratio has been found to a point of severity where the violation of variance homogeneity assumption seriously affects Type I error rates and effect size measures when sample sizes are unequal (see, Carol & Nordholm, 1975). In addition, the variance pattern is important when determining whether or not a linear or quadratic rule should be used to obtain hit rates when using PDA. As previously pointed out, if population variances are judged to be unequal, a quadratic rule is recommended to obtain an appropriate hit rate. The choice of linear versus quadratic classification is discussed more extensively by McLachlan (1992, pp. 132-137).

Three levels of total sample size were manipulated. Total sample size was initially varied at three levels, $\underline{N} = 40$, $\underline{N} = 100$, and $\underline{N} = 600$. Based on the Cohen (1988, p. 30) power charts, these sample sizes were sufficient to test the null hypothesis of no population mean difference with power equaling .80 at alpha equaling .05 in a directional test when the populations differ by $.80\sigma$, $.50\sigma$, and $.20\sigma$, respectively. However, using an iterative procedure, we found that the largest $\underline{N}$ needed was 300 because $\underline{N}$ sizes greater than 300 revealed no change in the sample estimates of $\underline{I}$. Thus the final three sample sizes used in this study were 40, 100, and 300.

For each level of $\underline{N}$, three patterns of group or $\underline{n}$ sizes were used. For $\underline{N} = 40$, sample size ratios of 20:20, 30:10 (where the larger $\underline{n}$ was associated with the smaller variance), and 10:30 (where the smaller $\underline{n}$ was associated with the smaller variance) were used. For $\underline{N} = 100$, $\underline{n}$ ratios were 50:50, 75:25, and 25:75, and for $\underline{N} = 300$, $\underline{n}$ ratios were 150:150, 225:75, and 75:225. Moreover, considering equal and unequal $\underline{n}$ ratios in combination with unequal variance patterns was viewed important to adequately describe the sampling characteristics of $\underline{I}$. Previous researchers (e.g., Glass, Peckham, & Sanders, 1972; Lix & Kesselman, 1998) have considered this joint condition to assess the robustness of common test statistics such as the $\underline{t}$ and $\underline{F}$ statistics.

Finally, two levels of population shape were considered: a normal and a skewed-leptokurtic or peaked (1.75, 3.75) distribution. The distribution shapes were identical for the two populations being compared. A third level of nonnormality (skewed-mesokurtic .75, 0) was initially considered but based on preliminary results this level was dropped as unnecessary. We thought considering only two distribution shapes was sufficient to obtain a good picture of the sampling properties of the indices under investigation.

Data Generation

We generated data to meet the above conditions using SAS IML (SAS Institute, 1990). Within each of the two populations, independent normally distributed observations $Z_{ij}$ ($i = 1 \ldots n_j$ and $j = 0$ or 1) were created using the SAS-RANNOR function. Using the Fleishman (1978) power transformation,

$$X_{ij} = a_{ij} + bZ_{ij} + cZ^2_{ij} + dZ^3_{ij} \quad , \tag{5}$$

the observations were transformed to reflect the target distribution shapes. That is, for normal distributions, a = 0, b = 1, c = 0, and d = 0. For the skewed-leptokurtic distributions, the constants were set in Equation 5 to: a = - .399, b = .930, c = .399, and d = -.036 (Fleishman, 1978). To generate data with the desired expected means and variances, each observation was transformed by multiplying it by $\sqrt{\sigma^2}$ and added to the desired population mean, $\mu_j$ ($Y_{ij} = \mu_j + X_{ij}\sqrt{\sigma_j^2}$). To arrive at the three levels of population separation (i.e., $\delta$ = .20, .50, and .80), differences in population means were standardized using the standard deviation of population one (or 1).

Data were then exported to SAS DISCRIM in order to obtain sample estimates of the population linear and quadratic hit rates and I values (here on in, sample estimates of I using linear and quadratic PDA are denoted as linear $\hat{I}$ and quadratic $\hat{I}$, respectively). Using the SAS

DISCRIM procedure, a predictive discriminant analysis was performed. First, to obtain sample estimates of population linear and quadratic hit rates we decided to set the prior probabilities to be equal (.50), which meant that in the population the probability of being in either of the two groups was .50. Second, an external leave-one-out (L-O-O) counting analysis was used to obtain hit rates based on a linear and a quadratic rule. As previously discussed by Huberty and Lowman (1998, p. 194), external analyses using an L-O-O counting estimator yield unbiased estimates of the true population values. To compute linear and quadratic $\hat{I}$ values, chance was defined by using the proportional chance criterion (recall that the formula under the proportional chance criterion is $\sum n_j q_j / N$). Subsequently, for this study $q_g$ or prior probabilities for both groups were .50.

Finally, the same data were exported to the SAS LOGISTIC in order to compute sample estimates of the population hit rates and $I$ values based on LRA (here on in, sample estimates of $I$ using LRA are denoted as logistic $\hat{I}$). A cutoff value of .50 was used to classify units into the target group (group 2). That is, .50 represents the modeled probability function for group 2 and was considered to be equal to that of group 1. In addition, similar to PDA, hit rates based on LRA are biased upward because model estimation and classification are done on the same sample. For PDA, this bias correction was achieved by implementing an external analysis (L-O-O counting estimation method). Conversely, for LRA, this external analysis technique wherein fitting the model with each observation left out was considered to computationally expensive (SAS Institute, 1997, p. 461). Instead of using a L-O-O counting approach, the SAS LOGISTIC procedure directly implements a less expensive one-step algebraic approximation for correcting the upward bias (SAS Institute, 1997, pp. 461-468). Logistic $\hat{I}$ values were computed similar to linear and quadratic $\hat{I}$.

Data Analyses and Evaluation

The four data conditions were manipulated for the present investigation in a completely crossed design. A total of 162 conditions were investigated: 3 levels of population separation, 3 variance ratios, 9 sample size levels (i.e., including equal and two unequal $\underline{n}$ ratios under each of three total sample sizes), and 2 distribution shapes. For each of these conditions 5,000 replications for each index were computed. In order to describe the distributional properties of each index, means, standard deviations, and three quantiles (i.e., 25th, 50th, and 75th percentiles, or $Q_1$, $Q_2$, and $Q_3$) were tabulated for each condition.

The accuracy or the degree of bias of each estimator was computed as the difference between the sample mean of $\hat{\underline{I}}$ over 5,000 replications and the true value of $\underline{I}$. Differences greater than +/- .30$\underline{I}$ (or, in other words, differences in excess of 30%) indicated severe bias. This 30% criterion was based on Bradley (1978) who recommended that a procedure might be considered robust to the violation of an assumption if the Type I error rate was within +/- .50$\alpha$. Bradley considered +/- .50$\alpha$ liberal and .10$\alpha$ conservative. Adopting Bradley's approach, we considered .50$\underline{I}$ to be too liberal and .10$\underline{I}$ to be too conservative, therefore we decided .30$\underline{I}$ was a reasonable criterion for bias. Finally, precision was computed as the standard deviation of the sampling distribution of $\hat{\underline{I}}$ under each condition. Box plots were also used to evaluate the precision of the estimators. Specifically, the inclusion of the median of $\hat{\underline{I}}$ at one level of population separation within the hinges (25th and 75th percentiles) of adjacent levels of population separation was viewed as unacceptable.

Determining Population Values of I

Based on Equation 3, a theoretical value for $\underline{I}$ can be determined assuming variance homogeneity and normality. Using the proportional chance criterion to determine the chance

hit rate as $H_e = \sum (q_g n_g) / N$, and the true hit rate determined as

$$\underline{H}_o = \underline{\phi}(\Delta/2) \ , \tag{6}$$

where $\underline{\phi}$ is the standard normal distribution function and $\Delta$ is the positive square root of the population Mahalanobis distance index (Huberty, 1994, p. 84), $\underline{I}$ can be computed as:

$$\underline{I} = [\underline{\phi}(\Delta/2) - \sum (\underline{q_g n_g}) \underline{N}^{-1}] / [1 - \sum (\underline{q_g n_g}) \underline{N}^{-1}] \ . \tag{7}$$

In the univariate case with normal distributions and equal variances, Mahalanobis distance $\Delta$ equals the standardized mean difference $\delta$. Furthermore, in the two-population comparison case, where the probability of being in population 1 is .50 and in population 2 is .50 (e.g., in a randomized experiment), the chance classification is simply .50 ($\sum (\underline{q_g n_g}) \underline{N}^{-1} = .50$). Then for different values of population separation, $\delta$, $\underline{I}$ can be computed. For example, when $\delta = .20$, $\phi(.2/2) = .539$ (see Equation 6) and $\underline{I} = .080$ (see Equation 7). Similarly for $\delta = .50$ and $\delta = .80$, $\underline{I}$ equals .197 and .311, respectively. Finally, under the optimal conditions of variance homogeneity and normality, theoretical across-group hit rates, $\underline{H}_o$, determined from LRA are identical to those using both linear and quadratic PDA.

When population variances are unequal and/or when population distributions are nonnormal, the aforementioned procedure for computing the true $\underline{H}_o$ is not so straightforward. To obtain $\underline{I}$ values under variance heterogeneity, it was necessary to determine the value of $\underline{I}$ in the population for each heterogeneous variance pattern under study (i.e., 1:4 and 1:8). For any given unequal variance pattern (e.g., 1:4), both linear PDA and LRA methods will provide identical values of $\underline{I}$. On the other hand, $\underline{I}$ values based on quadratic PDA are different for each variance pattern. In fact, as population variance patterns become more extreme, quadratic PDA will maximize the across-group hit rate, and in turn, population $\underline{I}$ values become larger. In order to obtain $\underline{I}$ values under variance heterogeneity and nonnormality, we generated total sample

sizes of 2,000,000 for each variance pattern/distribution shape of interest here and treated them as different "populations." Then for each "population" we computed the I index using the linear and quadratic PDA and LRA methods.

To check the accuracy of our empirically generated "population" values, we computed the values of I under normality and equal variance and compared them with the theoretical values derived using Equations 6 and 7. Table 1 shows the population I values for three variance patterns under data normality, nonnormality, and three levels of population separation. When population variances were equal and distributions normal, the empirically derived I values were almost identical to the theoretical values. We used these empirically derived values of I to evaluate the sample estimates of I under a variety of conditions.

To interpret the relative size of I across the data conditions, Table 1 reveals a general change in the size of I, depending on the variance pattern and distribution shape. For example, as population variance patterns become more extreme, I values derived from both linear PDA and LRA decrease but remain the same, while I values derived from a quadratic PDA tend to increase. The change in size is more obvious for larger effect sizes (.80) than for smaller effect sizes (.20). This makes sense because quadratic PDA maximizes the across-group hit rate under variance heterogeneity. Finally, under nonnormal population distributions, I values are in general smaller than those under normal distributions.

## Results

### Distributional Properties of I under Normal Population Distributions

Equal Variances. Table 2 contains results pertaining to the accuracy and precision of linear, quadratic, and logistic $\hat{I}$ under equal population variances (1:1) and data normality (0,0). Values in bold identify those conditions where the bias exceeded our criterion of +/- .3I. Under

these optimal conditions none of the indices were severely biased, except when both population separations were small ($\delta$ = .20) and sample sizes were small ($\underline{N}$ = 40). That is, when $\delta$ = .20 and $\underline{N}$ = 40 (and our cutoff was +/- .025), linear $\hat{\underline{I}}$ underestimated the parameter by .052, quadratic $\hat{\underline{I}}$ underestimated the parameter by .056, and logistic $\hat{\underline{I}}$ overestimated the parameter by .049 for equal $\underline{n}$ sizes. A similar pattern was found when $\underline{n}$ sizes were unequal, thus indicating consistency across $\underline{n}$ ratios. Table 2 also indicates that linear and quadratic PDA slightly exceeded our criterion of unacceptable bias when $\underline{N}$ sizes were moderate ($\underline{N}$ = 100) for equal $\underline{n}$ sizes.

Although the LRA approach resulted in the best precision, all indices varied greatly from the parameter value. The precision of all indices improved when $\underline{N}$ sizes were large ($\underline{N}$ = 300), as shown in Table 2. In addition, Figure 1 graphically demonstrates the precision of the parameter estimation by presenting the three point summaries ($Q_1$, $Q_2$, $Q_3$) of the sampling distributions for each index when $\underline{N}$ = 300 (equal $\underline{n}$ sizes). When $\underline{N}$ = 300, the median of each $\hat{\underline{I}}$ index at one level of population separation was not captured within the hinges of adjacent levels of population separation. However, when $\underline{N}$ sizes were smaller, this was not the case. For example, when $\underline{N}$ = 40 and 100 the median of each $\hat{\underline{I}}$ was included within the hinges of adjacent levels of population separation (this result is not shown in Figure 1 but is available in supplementary figures). Thus, unless the sample size was large, there was considerable overlap among each index's sampling distributions.

Heterogeneous variances. Table 3 presents the results for each index when populations variances were moderately heterogeneous (1:4) [1] but population distributions were normal. When $\underline{n}$ sizes were equal, linear and logistic $\hat{\underline{I}}$ provided estimates of the parameter that were generally within our acceptable criterion, except when both population separations and sample

sizes were small ($\delta$ = .20 and $\underline{N}$ = 40). That is, Table 3 shows when $\delta$ = .20 and $\underline{N}$ = 40 (and our cutoff was +/- .019), linear $\hat{\underline{I}}$ underestimated the parameter by .023 and logistic $\hat{\underline{I}}$ overestimated the parameter by .068. Quadratic $\hat{\underline{I}}$ on the other hand provided estimates within our acceptable criterion when $\underline{n}$ sizes were equal across all levels of population separation and $\underline{N}$ size. Moreover, the results in Table 3 were consistent with those found under more extreme variance heterogeneity (1:8).

With unequal $\underline{n}$ sizes and moderate variance heterogeneity (1:4), both linear and logistic $\hat{\underline{I}}$ provided estimates of the parameter that exceeded our criterion for bias only when group separation was small ($\delta$ = .20). Specifically, Table 3 shows with small population separation, linear $\hat{\underline{I}}$ underestimated the parameter when the group with the smaller $\underline{n}$ had the smaller variance, while logistic $\hat{\underline{I}}$ overestimated the parameter when the group with the smaller $\underline{n}$ had the larger variance. With larger group separation ($\delta$ = .5 or .8), both linear and logistic $\hat{\underline{I}}$ provided acceptable estimates of the parameter. Quadratic $\hat{\underline{I}}$ on the other hand consistently over- or underestimated the parameter across all levels of population separation and $\underline{N}$ sizes. Furthermore, for extreme variance heterogeneity (1:8), none of the indices provided estimates of the parameter that was within our criterion for bias when $\underline{n}$ sizes were unequal. Thus indicating that $\underline{I}$ derived from either PDA or LRA leads to severely biased estimates of the parameter when variance heterogeneity is extreme and $\underline{n}$ sizes are unequal.

The precision of each index slightly improved when $\underline{N}$ sizes were large ($\underline{N}$ = 300), as shown in Table 3. Figure 2 further shows when variance heterogeneity was moderate (1:4) and $\underline{N}$ = 300, the medians of both linear and logistic $\hat{\underline{I}}$ were just captured within the hinges of adjacent levels of effect size. Thus indicating some overlap among each index's sampling distributions. Under extreme variance heterogeneity (1:8), the precision of linear and logistic $\hat{\underline{I}}$

did not improve at all under large N. Furthermore, with quadratic $\hat{I}$, the three levels of group separation under moderate variance heterogeneity (1:4) resulted in I values that differed only slightly, .463, .470, and .480 for small, medium, and large group separation, respectively. In other words, the sampling distributions of I derived from quadratic PDA were almost identical for all three levels of effect size. Consequently, when variances differ it is almost impossible to distinguish among the three levels of group separation studied using quadratic PDA.

Distributional Properties of I under Nonnormal Population Distributions

Equal variances. Table 4 summarizes the results when population variances were equal (1:1) and population distributions were nonnormal (1.75, 3.75). When n sizes were equal, all indices provided acceptable estimates of the parameter except when both population separations and sample sizes were small. That is, when $\delta = .20$ and $N = 40$ (and our cutoff was +/- .020), linear $\hat{I}$ underestimated the parameter by .036, quadratic $\hat{I}$ underestimated the parameter by .020, and logistic $\hat{I}$ overestimated the parameter by .038. With unequal n sizes, none of the indices provided an estimate of the population value that was within our criterion for bias. Specifically, Table 4 shows that all estimates overestimated the population value when group 1 (or the target group in the LRA case) had the larger n, and tended to underestimate the population value when group 2 had the larger n. This is in contrast to what was found for unequal n sizes under data normality (see Table 2).

The precision of the each index under nonnormal population distributions was typically less than when the population distributions were normal. For example, Table 2 shows when data were normal and n sizes were equal, the precision of each index was .069, .061, .049 for linear, quadratic, and logistic $\hat{I}$, respectively for small population separations ($\delta = .20$) and large sample sizes ($N = 300$). For these same conditions but nonnormal distributions, as shown in Table 4, the

precision was .091, .072, and .053, respectively. Figure 3 further demonstrates the small reduction in precision under nonnormality when sample sizes were large. That is, when N = 300, the medians of each $\hat{I}$ index were just captured within the hinges of adjacent levels of effect size. However, compared to data normality, when N = 300, the median of each $\hat{I}$ index was not captured within the hinges of adjacent levels of effect sizes (see Figure 1).

Heterogeneous variances. When population variances were heterogeneous and distributions were nonnormal, results were found to be similar to those under equal variances (1:1) and nonnormality. Table 5 presents the results for moderate variance heterogeneity (1:4) and nonnormal populations. When n sizes were equal, all indices provided estimates of the parameter that were within our criterion for bias, except when both population separations were small ($\delta$ = .20) and when sample sizes were small (N = 40). However, when n sizes were unequal, none of the indices provide an estimate of the parameter that was within our criterion for bias. Furthermore, the degree of accuracy shown in Table 5 was similar to that found under extreme variance heterogeneity (1:8).

The precision of each index again did not greatly improve under variance heterogeneity and nonnormal distributions when sample sizes were large. For example, Table 4 shows when data were nonnormal and variances were equal, the precision of each index was .091, .072, and .053 for linear, quadratic, and logistic $\hat{I}$, respectively for equal n sizes, small population separations ($\delta$ = .20), and large sample sizes (N = 300). For these same conditions but moderately unequal variances, as shown in Table 5, the precision of each index was .124, .079, and .056. Figure 4 further demonstrates the small reduction in the precision under nonnormality when variance heterogeneity was moderate (1:4) and sample sizes were large. That is, when N = 300, the median of each $\hat{I}$ index was just captured within the hinges of adjacent levels of effect

size. Compared to the equal variances condition and data nonnormality, when $\underline{N}$ = 300, the median of each $\hat{\underline{I}}$ index was also captured within the hinges of adjacent effect sizes (see Figure 3). Furthermore, under extreme variance heterogeneity (1:8) and nonnormality, the precision was even less than under moderate variance heterogeneity.

## Discussion

None of the indices studied provided and adequate estimate of effect size for all of the conditions studied. The usefulness of each of the indices depended on the population characteristics. Under the optimal conditions of equal population variances and data normality, both linear and quadratic PDA and LRA provided accurate estimates of $\underline{I}$, except when population separations and sample sizes were jointly small ($\delta$ = .20 and $\underline{N}$ = 40). These results were consistent across all $\underline{n}$ ratios. In addition, linear and quadratic PDA methods also led to unacceptable bias under small population separations ($\delta$ = .20) and moderate sample sizes ($\underline{N}$ = 100). Furthermore, the precision of all indices under optimal conditions was good only under large sample sizes.

When variances were heterogeneous, disparities between PDA and LRA depended on the $\underline{n}$ ratio. When $\underline{n}$ sizes were equal, both linear PDA and LRA provided accurate estimates of $\underline{I}$, except when both population separation was small ($\delta$ = .20) and the total sample size was small ($\underline{N}$ = 40). Quadratic PDA, on the other hand, provided an accurate estimate of the parameter regardless of the degree of population separation and sample size. Conversely, when $\underline{n}$ sizes were unequal, linear PDA and LRA led to severely biased estimates only when variance heterogeneity was moderate (1:4) and when population separation was small ($\delta$ = .20). The LRA method overestimated the population value when the group with the smaller $\underline{n}$ had the larger variance. Similarly, quadratic PDA consistently overestimated the parameter, but did so across

all conditions. Under extreme variance heterogeneity (1:8), none of the indices performed well when n sizes were unequal.

The unacceptable bias found when both variances and n sizes were unequal may have resulted from using equal prior probabilities (for the PDA method). Under heterogeneous variances, if the n sizes do not reflect the sizes of the two populations, greater bias may result. Similarly, the cutoff value should match the size of the target population when using the LRA method to derive I.

In terms of precision, neither PDA nor LRA estimates of I were very precise under variance heterogeneity, even when sample sizes were large. The lack of stability presented by quadratic PDA may partly be due to the inability to differentiate between small, medium, and large population values of I under variance heterogeneity. When population variances are unequal, the goal of quadratic PDA is to maximize the across-group hit rate. As variances become more heterogeneous, the relative sizes of the theoretical hit rates using quadratic PDA become less distinguishable, rendering the relative sizes of I also indistinguishable. This poses as a major limitation of quadratic PDA to derive I under variance heterogeneity.

Nonnormality impacted all three methods of computing I. When n sizes were equal, all three methods of computing I provided an adequate estimate of the parameter except when population separation and sample size were jointly small. This was similar to the result under data normality. However, precision, in general, was less than what resulted under data normality. Unlike under data normality, the variability of sample estimates of I did not decrease greatly with large sample sizes. Conversely, when n sizes were unequal all three methods of deriving I led to severely biased estimates across all population separations and sample sizes. The bias was upward when the larger n was associated with the smaller variance and downward

when the larger n was associated with the larger variance. Again, this may be largely due to the fact that equal priors were used (in PDA) or a .50 cutoff value was used (in LRA) across all conditions.

Finally, under the joint occurrence of heterogeneous variances and nonnormality, the results were similar to those found under equal population variances and nonnormal distributions. When n sizes were equal, none of the indices were severely biased regardless of the distribution shape. However, when n sizes were unequal, all of the indices were severely biased. As in the case of data normality, the discrepancy between the size of the priors and n ratio may have been responsible for the inaccuracy found under nonnormality.

Practical Implications and Recommendations

Under optimal conditions, either PDA or LRA are acceptable methods for deriving I, provided population separation and sample size are not jointly small (i.e., $\delta = .20$ and $N = 40$). When variances are heterogeneous, LRA is more practical compared to PDA. Although Huberty and Lowman (2000) recommended the quadratic rule when computing I when variances are judged to be heterogeneous, there are two major limitations of using this procedure. First, as shown in the present study, when variance patterns become more extreme, the true values of I based on quadratic PDA become less differentiated. Second, given the difficulties associated with statistical tests for variance equality (e.g., Box test) used under nonnormality, it may be difficult to determine exactly when the quadratic rule should be used. LRA, on the other hand, does not require a test of variance equality, the values of I are differentiated for all variance patterns and is easy to compute, thus representing a more practical method to assess group overlap.

Given the practical limitations of using quadratic PDA, we recommend that LRA be used to estimate I under variance heterogeneity if group or n sizes are equal. When group sizes are unequal, we also recommend LRA to compute I over quadratic PDA only if variance heterogeneity is moderate (1:4) and population separation is not small (i.e., δ values .50 or greater). When distributions are nonnormal, we recommend LRA regardless of the variance pattern, provided n sizes are equal. Finally, we should point out that the precision or stability of sample estimates was in general somewhat better only under large sample sizes (N = 300). Therefore, for best performance, I derived from LRA should be used if sample sizes are large. Moreover, if one chooses to use LRA under the data conditions stated above, the following intervals are suggested:

< .08 is small
.11 to .15 is medium
> .20 is large.

As demonstrated in Table 1, these intervals (including gaps between them) were created because small, medium, and large I values (based on δ) slightly shift downward when variances become more heterogeneous and when distribution shapes are nonnormal.

Limitations

There are three aspects of this study that may limit the generalizability of the findings. First, we selected and manipulated a limited number of data conditions, and so the findings can only be generalized to the specific data conditions and levels used in the present study. Although the specific levels under each condition did provide sufficient information to adequately describe the sampling properties of each index, future research might consider additional levels in order to obtain a more comprehensive picture.

Second, we investigated only two-group univariate mean comparisons. Thus the conclusions drawn from this study pertain exclusively to the two-group comparison context. However we suspect similar and perhaps more extreme problems would arise in more complex analyses (e.g., multiple groups, multiple outcomes). Future research might consider investigating the performance of I, perhaps using polytomous logistic regression, when the number of groups being compared is greater than two.

Third, only equal population based priors (.50 and .50) were used in the context of PDA, and only a .50 cutoff value was used in the context of LRA. Future research may examine the effect of using priors based on sample proportions instead of known population sizes (i.e., assuming proportional sampling). This would be important because when group sizes were unequal, I in general was severely biased under variance heterogeneity and nonnormality. We believe that in practice many researchers may be inclined to simply use sample proportions if no knowledge of population sizes is available. Unless a proportional sampling procedure was used, this may lead to improper hit rates and in turn improper I values when variances are unequal.

Practical Limitations of using I as a Measure of Effect Size

One of the major shortcomings of using the I index, particularly derived from PDA, is that, depending on the ratio of the variances and distribution shape, a different theoretical value of I will be obtained. In the case of quadratic PDA, I values are less differentiated in terms of small, medium, and large as variance patterns become more extreme. Likewise with linear PDA and LRA, I values tend to attenuate as variances become more extreme. Because most of the data that are gathered in the social sciences for instance manifest conditions that are less than ideal, researchers cannot attempt to make stringent qualitative judgments based on sample estimates of I regardless of which method is used to compute I. Thus, under less than ideal data

conditions, we stress that any I index computed from a given data set a must be interpreted with caution, even when using the intervals suggested in this study.

Conclusion

Under optimal data conditions, the researcher has the luxury of using either PDA or LRA to derive acceptable estimates of I, except when both sample size and population separation are small. However, this joint condition may be avoided if the researcher performs a power analysis a priori to maintain the proper sample size to detect an anticipated effect size in the population. Furthermore, under heterogeneous variances, quadratic PDA is not recommend because of an inability to differentiate between levels of effect size, making qualitative interpretation ditficult. LRA on the other hand does not require a test of variance equality and logistic regression-based hit rates can be easily obtained from popular statistical software packages (e.g., SPSS and SAS), making LRA the more practical derivation method. For optimal performance, however, I should be used under large sample sizes (N = 300) because of improved precision (and accuracy).

To conclude, we recommend I derived from LRA unless n sizes are unequal, in which case I derived from LRA is acceptable only when variance heterogeneity is moderate (1:4) and population separation is not small. We believe this seems reasonable because small population separations are typically not desirable and more extreme variance heterogeneity (e.g., 1:8) conditions are atypical in social science research. Finally, under nonnormal population distributions, I derived from LRA is recommended regardless of the variance pattern provided n sizes are equal. Hence, it appears the n ratio is a critical factor as to whether one can use LRA to compute I when variances are heterogeneous and/or under nonnormality. However, based on the conclusions drawn form this study, if the researcher can a priori maintain an equal n ratio when the size of the two populations are equal, then the use of I derived from LRA can be efficaciously used.

References

Alf, E., & Abrahams, N. M. (1968). Relationship between percent of overlap and measures of correlation. Educational and Psychological Measurement, 28, 779-792.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Carroll, R. M., & Nordholm, L. A. (1975). A sampling characteristic of Kelly's $\varepsilon^2$ and Hays' $\omega^2$. Educational and Psychological Measurement, 35, 541-554.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1988). Statistical power analysis of the behavioral sciences (2nd ed.). New York: Academic Press.

Cox, D. R., & Snell, E. J. (1989). The analysis of binary data (2nd ed.). London: Chapman and Hall.

Dattalo, P. (1994). A comparison of discriminant analysis and logistic regression. Journal of Social Service Research, 19, 121-144.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association, 70, 892-898.

Elster, R. S., & Dunnette, M. D. (1971). The robustness of Tilton's measure of overlap. Educational and Psychological Measurement, 31, 685-697.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard. Theory & Psychology, 5, 75-98.

Fan, X., & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. Journal of Experimental Education, 67, 265-286.

Fleishman, A. (1978). A method for simulating nonnormal distributions. Psychometrika, 43, 521-531.

Fowler, R. L. (1988). Estimating the standardized mean difference in intervention studies. Journal of Educational Statistics, 13, 337-350.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Greenwald, A. G., Gonzalez, R., Harris, R. L., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated. Psychophysiology, 33, 175-183.

Harwell, M. R., Rubenstein, E. N., Hays, W: S., & Olds, C. C. (1992). Summarizing Monte Carlo research in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17, 315-339.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6, 107-128.

Huberty, C. J (1994). Applied discriminant analysis. New York: Wiley.

Huberty, C. J, & Holmes, S. E. (1983). Two-group comparisons and univariate classification. Educational and Psychological Measurement, 43, 15-26.

Huberty, C. J, & Lowman, L. L. (1998). Discriminant analysis in higher education research. In J.C. Smart (Ed.), Higher education: Handbook of theory and research (pp. 181-234). New York: Agathon.

Huberty, C. J, & Lowman, L. L. (2000). Group overlap as a basis for effect size. Educational and Psychological Measurement, 60, 543-563.

Huberty, C. J, & Pike, C. J. (1999). On some history regarding statistical testing. In B. Thompson (Ed.), Advances in social science methodology, Vol. 5. (pp. 1-22). Greenwich, CT: JAI Press.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Levy, P. (1967). Substantive significance of significant differences between two groups. Psychological Bulletin, 67, 37-40.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Test of location equality under heteroscedasticity and nonnormality. Educational and Psychological Measurement, 58, 409-429.

McLachlan, G. J. (1992). Discriminant analysis and statistical pattern recognition. New York: Wiley.

Meshbane, A., & Morris, J. D. (1996, April). Predictive discriminant analysis verses logistic regression in two-group classification problems. Paper presented at the annual meeting of the American Educational Research Association, New York.

Neter, J., Wasserman, W., & Kutner, M. H. (1989). Applied linear regression models (2nd ed.). Boston: Irwin.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. Contemporary Educational Psychology, 25, 241-286.

Oshima, T. C., & Algina, J. (1992). A SAS program for testing the hypothesis of the equal means under heteroscedasticity: James's second-order test. Educational and Psychological Measurement, 52, 117-118.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association, 73, 699-705.

Richardson, J. T. E. (1996). Measures of effect size. Behavior Research Methods, Instruments, & Computers, 28, 12-22.

SAS Institute Inc. (1990). SAS / IML software: Usage and reference, version 6 (1st ed.). Cary, NC: Author.

SAS Institute Inc. (1997). SAS / STAT software: Changes and enhancements through release 6.12. Cary, NC: Author.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

Strube, M. J. (1988). Some comments on the use of magnitude-of-effect estimates. Journal of Counseling Psychology, 35, 342-345.

Thompson, B. (1999a). Statistical significant tests, effect size reporting, and the vain pursuit of pseudo-objectivity. Theory & Psychology, 9, 191-196.

Thompson, B. (1999b). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to changing practices. Journal of Psychology, 133, 133-140.

Tilton, J. W. (1937). The measurement of overlapping. Journal of Educational Psychology, 28, 656-662.

Wilcox, R. R. (1987). New designs in analysis of variance. Annual Review of Psychology, 38, 29-60.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

## Authors' Note

Brian Hess is now at Psychological Assessment Resources, Odessa, FL; Stephen Olejnik and Carl J Huberty, Department of Educational Psychology, University of Georgia.

We would like to thank Dr. Joseph Wisenbaker for his assistance with the computer program and Drs. David A. Payne and Hubert Chen for their constructive feedback on an earlier draft of this manuscript.

Correspondence concerning this article should be addressed to Brian Hess, 12711 Arbor Isle Drive, Temple Terrace, FL 33637. Electronic mail may be sent to bhess@parinc.com.

## Footnotes

[1] Due to space restrictions, only a representative sample of conditions are reported in this article. Specifically, supplementary tables containing results under levels of extreme variance heterogeneity (1:8) can be obtained from the first author.

Table 1

Empirically Derived Values of I Based on PDA and LRA under Normal (0, 0) and Nonnormal (1.75, 3.75) Population Distributions

| | $\delta = .20$ | | | |
|---|---|---|---|---|
| | Linear PDA / LRA | | Quadratic PDA | |
| Variance Pattern | Normal | Nonnormal | Normal | Nonnormal |
| 1:1 | .080 | .065 | .080 | .065 |
| 1:4 | .060 | .048 | .324 | .298 |
| 1:8 | .053 | .044 | .463 | .499 |

| | $\delta = .50$ | | | |
|---|---|---|---|---|
| | Linear PDA / LRA | | Quadratic PDA | |
| | Normal | Nonnormal | Normal | Nonnormal |
| 1:1 | .197 | .166 | .197 | .166 |
| 1:4 | .148 | .117 | .341 | .147 |
| 1:8 | .133 | .105 | .470 | .430 |

| | $\delta = .80$ | | | |
|---|---|---|---|---|
| | Linear PDA / LRA | | Quadratic PDA | |
| | Normal | Nonnormal | Normal | Nonnormal |
| 1:1 | .311 | .269 | .311 | .269 |
| 1:4 | .235 | .182 | .367 | .187 |
| 1:8 | .212 | .160 | .480 | .278 |

Table 2

Accuracy and Precision of PDA and LRA Estimates of I under Equal Variances (1:1) and Normal Population Distributions (0, 0)

| | | Population Separation | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | δ = .20 | | | | | | δ = .50 | | | | | | δ = .80 | | | | | |
| N size | n ratio | Linear PDA (I = .080) | | Quadratic PDA (I = .080) | | Logistic (I = .080) | | Linear PDA (I = .197) | | Quadratic PDA (I = .197) | | Logistic (I = .197) | | Linear PDA (I = .311) | | Quadratic PDA (I = .311) | | Logistic (I = .311) | |
| | | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P |
| 40 | 20:20 | **-.052** | .287 | **-.056** | .226 | **.049** | .130 | -.029 | .223 | -.048 | .205 | .006 | .148 | -.010 | .166 | -.028 | .172 | .002 | .151 |
| | 30:10 | -.021 | .259 | -.021 | .247 | **.057** | .136 | -.026 | .235 | -.042 | .237 | .016 | .158 | -.014 | .187 | -.044 | .207 | .004 | .163 |
| | 10:30 | **-.027** | .269 | **-.025** | .250 | **.057** | .286 | -.028 | .233 | -.044 | .235 | .014 | .159 | -.011 | .186 | -.032 | .207 | .006 | .165 |
| 100 | 50:50 | **-.025** | .187 | **-.025** | .140 | .013 | .087 | -.003 | .110 | -.016 | .113 | .003 | .098 | -.011 | .095 | -.066 | .097 | .002 | .095 |
| | 75:25 | -.021 | .184 | -.024 | .195 | .020 | .092 | -.008 | .130 | -.021 | .159 | .001 | .108 | .000 | .104 | -.007 | .124 | .004 | .104 |
| | 25:75 | -.019 | .183 | -.023 | .189 | .020 | .092 | -.006 | .127 | -.028 | .159 | .002 | .107 | -.003 | .102 | -.010 | .123 | .000 | .103 |
| 300 | 150:150 | -.003 | .069 | -.011 | .061 | .001 | .049 | -.001 | .049 | -.002 | .049 | .000 | .005 | .000 | .048 | -.001 | .048 | .000 | .048 |
| | 225:75 | -.004 | .077 | -.012 | .135 | .002 | .052 | .000 | .055 | -.001 | .041 | .000 | .055 | -.022 | .053 | -.002 | .061 | .001 | .053 |
| | 75:225 | -.008 | .092 | -.012 | .135 | .000 | .054 | .001 | .054 | -.001 | .049 | .001 | .054 | -.001 | .052 | -.002 | .061 | .000 | .052 |

Note. A = accuracy or the difference between the sample mean of $\hat{I}$ and I; P = precision or the variance of $\hat{I}$. Values in bold identify those conditions where the bias exceeded our criterion of +/- .3I.

Table 3

Accuracy and Precision of PDA and LRA Estimates of I under Variance Heterogeneity (1:4) and Normal Population Distributions

(0,0)

| | | Population Separation | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | δ = .20 | | | | | | δ = .50 | | | | | | δ = .80 | | | | | |
| N size | n ratio | Linear PDA (I = .060) | | Quadratic PDA (I = .324) | | Logistic (I = .060) | | Linear PDA (I = .148) | | Quadratic PDA (I = .341) | | Logistic (I = .148) | | Linear PDA (I = .235) | | Quadratic PDA (I = .367) | | Logistic (I = .235) | |
| | | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P |
| 40 | 20:20 | **-.023** | .283 | -.028 | .153 | **.068** | .135 | -.042 | .270 | -.029 | .152 | .022 | .150 | -.036 | .240 | -.027 | .146 | -.001 | .158 |
| | 30:10 | **.051** | .275 | **.120** | .165 | **.122** | .162 | .021 | .273 | **.116** | .162 | .044 | .176 | .012 | .270 | **.114** | .153 | .049 | .188 |
| | 10:30 | **-.050** | .278 | **-.172** | .187 | **.044** | .124 | **-.058** | .247 | **-.170** | .179 | .002 | .137 | -.061 | .208 | **-.163** | .179 | -.029 | .145 |
| 100 | 50:50 | **-.023** | .207 | -.007 | .090 | .003 | .099 | -.015 | .162 | -.010 | .089 | .004 | .103 | -.005 | .121 | -.010 | .089 | -.002 | .108 |
| | 75:25 | .01[illegible] | .194 | **.152** | .090 | **.066** | .110 | .016 | .184 | **.148** | .086 | .038 | .126 | .027 | .165 | **.145** | .083 | .015 | .133 |
| | 25:75 | **-.035** | .199 | **-.167** | .110 | .016 | .080 | -.037 | .152 | **-.165** | .111 | -.021 | .092 | -.042 | .104 | **-.161** | .109 | -.037 | .095 |
| 300 | 150:150 | -.012 | .129 | -.004 | .051 | .007 | .057 | -.001 | .070 | -.004 | .052 | .000 | .063 | -.002 | .062 | -.003 | .052 | -.003 | .061 |
| | 225:75 | .013 | .121 | **.165** | .049 | **.030** | .071 | .023 | .097 | **.158** | .048 | .022 | .083 | .037 | .086 | **.154** | .048 | .027 | .080 |
| | 75:225 | **-.023** | .126 | **-.163** | .061 | -.003 | .051 | -.024 | .057 | **-.164** | .062 | -.023 | .057 | -.034 | .071 | -.086 | .051 | -.020 | .062 |

Note. A = Accuracy or the difference between the sample mean of $\hat{I}$ and I; P = Precision or the variance of $\hat{I}$. Values in bold identify those conditions where the bias exceeded our criterion of +/- .3I.



33

Table 4

Accuracy and Precision of PDA and LRA Estimates of I under Equal Variances (1:1) and Nonnormal Population Distributions (1.75, 3.75)

| | | Population Separation | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | δ = .20 | | | | | | δ = .50 | | | | | | δ = .80 | | | | | |
| N size | n ratio | Linear PDA (I = .065) | | Quadratic PDA (I = .065) | | Logistic (I = .065) | | Linear PDA (I = .166) | | Quadratic PDA (I = .166) | | Logistic (I = .166) | | Linear PDA (I = .269) | | Quadratic PDA (I = .269) | | Logistic (I = .269) | |
| | | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P |
| 40 | 20:20 | **-.036** | .272 | **-.020** | .191 | **.038** | .125 | -.023 | .229 | .020 | .177 | .022 | .151 | .002 | .170 | .031 | .171 | .035 | .169 |
| | 30:10 | **.051** | .299 | **.058** | .291 | **.104** | .181 | **.075** | .250 | **.098** | .196 | **.108** | .167 | **.104** | .163 | **.111** | .157 | **.111** | .145 |
| | 10:30 | **-.066** | .260 | **-.044** | .294 | **-.017** | .169 | **-.109** | .219 | **-.044** | .269 | -.034 | .190 | **-.102** | .212 | -.044 | .250 | -.011 | .240 |
| 100 | 50:50 | -.018 | .180 | -.004 | .124 | .019 | .083 | -.004 | .110 | .025 | .120 | .007 | .016 | .001 | .098 | .025 | .114 | .022 | .108 |
| | 75:25 | **.058** | .234 | **.048** | .246 | **.096** | .146 | **.117** | .117 | **.115** | .123 | **.116** | .099 | **.111** | .088 | **.112** | .100 | **.108** | .088 |
| | 25:75 | **-.080** | .176 | **-.048** | .256 | **-.051** | .123 | **-.126** | .130 | **-.062** | .212 | **-.087** | .129 | **-.106** | .120 | -.055 | .183 | -.044 | .150 |
| 300 | 150:150 | -.004 | .091 | .012 | .072 | .004 | .053 | .000 | .056 | .011 | .065 | .006 | .058 | .001 | .057 | .008 | .060 | .020 | .062 |
| | 225:75 | **.109** | .130 | **.078** | .178 | **.116** | .090 | **.124** | .055 | **.118** | .077 | **.119** | .053 | **.110** | .053 | **.110** | .059 | **.105** | .052 |
| | 75:225 | **-.118** | .105 | **-.075** | .222 | **-.105** | .081 | **-.125** | .065 | **-.091** | .145 | **-.103** | .076 | **-.108** | .068 | **-.088** | .105 | -.053 | .085 |

Note. A = Accuracy or the difference between the sample mean of $\hat{I}$ and I; P = Precision or the variance of $\hat{I}$. Values in bold identify those conditions where the bias exceeded our criterion of +/- .3I.

Table 5

Accuracy and Precision of PDA and LRA Estimates of I under Variance Heterogeneity (1:4) and Nonnormal Population Distributions (1.75, 3.75)

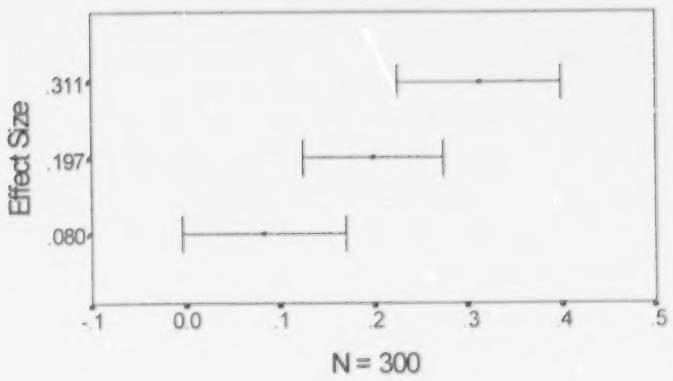
| | | Population Separation | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | δ = .20 | | | | | | δ = .50 | | | | | | δ = .80 | | | | | |
| N size | n ratio | Linear PDA (I = .048) | | Quadratic PDA (I = .298) | | Logistic (I = .048) | | Linear PDA (I = .117) | | Quadratic PDA (I = .147) | | Logistic (I = .117) | | Linear PDA (I = .182) | | Quadratic PDA (I = .187) | | Logistic (I = .182) | |
| | | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P | A | P |
| 40 | 20:20 | **-.021** | .268 | **-.093** | .179 | **.063** | .124 | -.032 | .245 | .015 | .158 | -.033 | .132 | -.020 | .142 | -.014 | .103 | .007 | .042 |
| | 30:10 | **.056** | .304 | **.115** | .245 | **.130** | .198 | **-.051** | .295 | **.215** | .250 | **.050** | .194 | **.091** | .272 | **.191** | .253 | **.113** | .183 |
| | 10:30 | **-.079** | .279 | **-.178** | .233 | **-.015** | .166 | **-.193** | .227 | **-.192** | .210 | **-.140** | .152 | **-.154** | .195 | **-.248** | .191 | **-.108** | .154 |
| 100 | 50:50 | **-.021** | .198 | -.005 | .119 | .011 | .080 | -.011 | .142 | .022 | .081 | .002 | .091 | -.004 | .107 | -.002 | .086 | -.003 | .095 |
| | 75:25 | **.060** | .236 | **.203** | .140 | **.097** | .162 | **.106** | .204 | **.296** | .118 | **.118** | .141 | **.141** | .144 | **.266** | .109 | **.129** | .110 |
| | 25:75 | **-.098** | .199 | **-.218** | .175 | **-.045** | .132 | **-.148** | .141 | **-.263** | .122 | **-.213** | .120 | **-.151** | .098 | **-.276** | .108 | **-.133** | .100 |
| 300 | 150:150 | .013 | .124 | -.001 | .079 | .010 | .056 | -.001 | .061 | .001 | .045 | -.002 | .056 | .000 | .055 | -.001 | .047 | -.002 | .055 |
| | 225:75 | **.086** | .175 | **.240** | .053 | **.149** | .132 | **.142** | .092 | **.309** | .040 | **.134** | .076 | **.152** | .066 | **.282** | .041 | **.133** | .100 |
| | 75:225 | **-.170** | .136 | **-.239** | .127 | **-.049** | .096 | **-.150** | .052 | **-.304** | .061 | **-.202** | .056 | **-.154** | .084 | **-.282** | .061 | **-.141** | .058 |

Note. A = Accuracy or the difference between the sample mean of $\hat{I}$ and I; P = Precision or the variance of $\hat{I}$. Values in bold identify those conditions where the bias exceeded our criterion of +/- .3I.

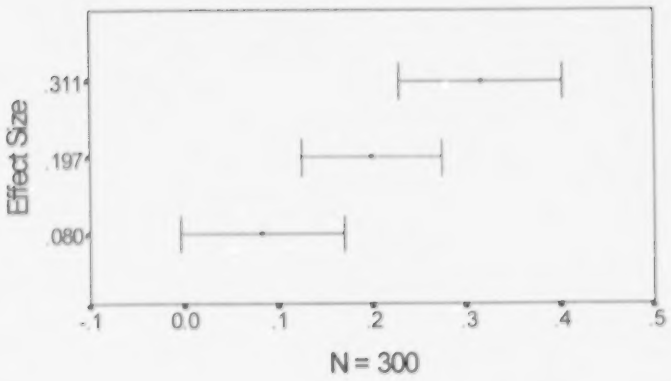
## Figure Caption

Figure 1. 3-Point Summary of Linear, Quadratic, and Logistic $\hat{I}$ under Optimal Conditions and Large Sample Sizes.

Linear PDA




Quadratic PDA




Logistic

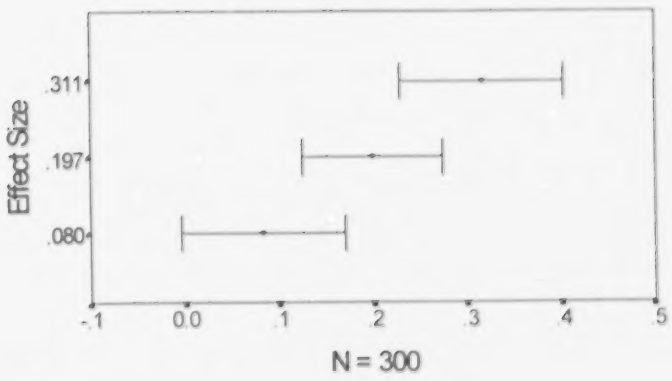
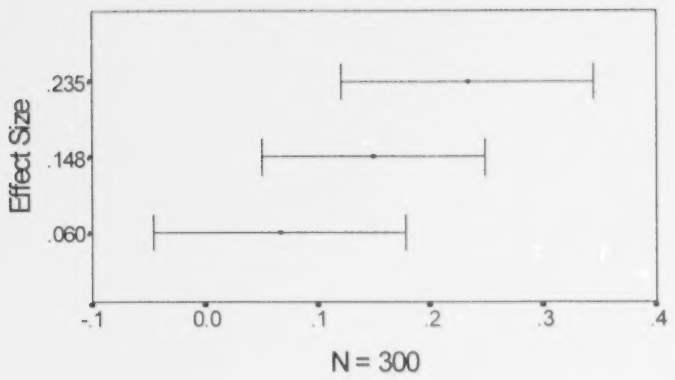


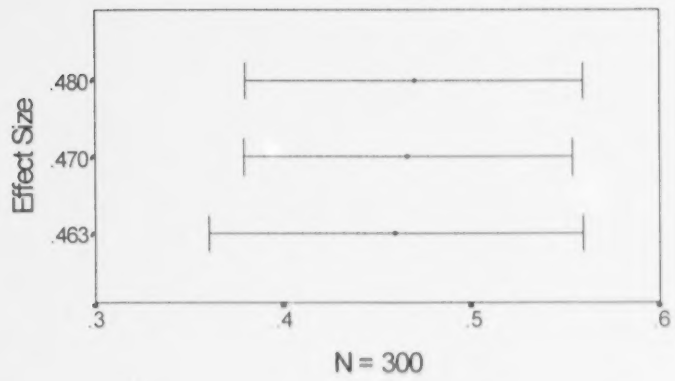
Figure Caption

Figure 2. 3-Point Summary of Linear, Quadratic, and Logistic $\hat{I}$ under Variance Heterogeneity (1:4), Normality (0, 0), and Large Sample Sizes.

Linear PDA




Quadratic PDA




Logistic

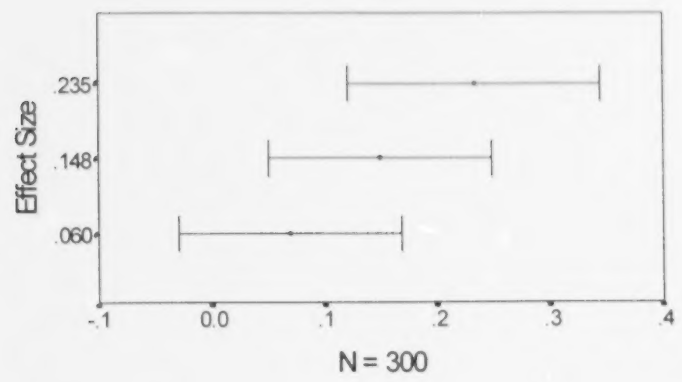
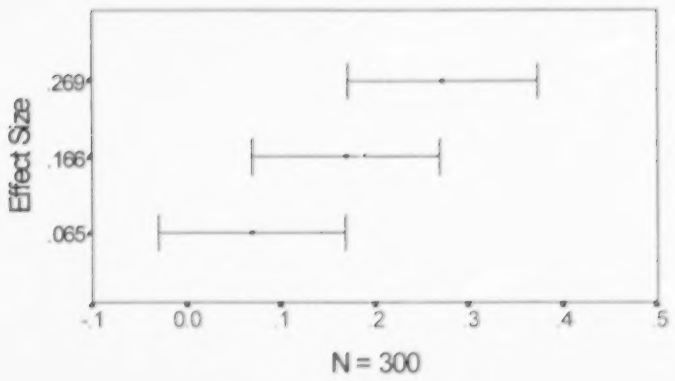


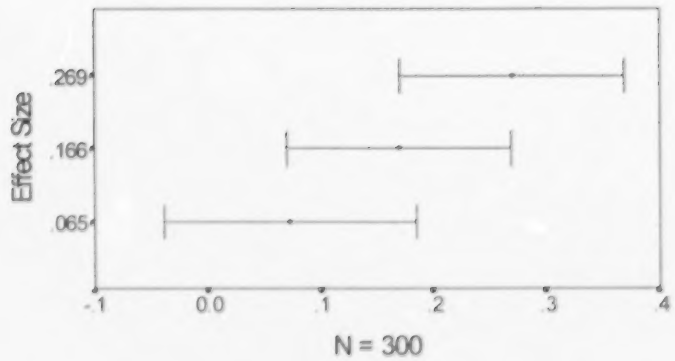
Figure Caption

Figure 3. 3-Point Summary of Linear, Quadratic, and Logistic $\hat{I}$ under Equal Variances (1:1), Nonnormality (1.75, 3.75), and Large Sample Sizes.

Linear PDA




Quadratic PDA




Logistic

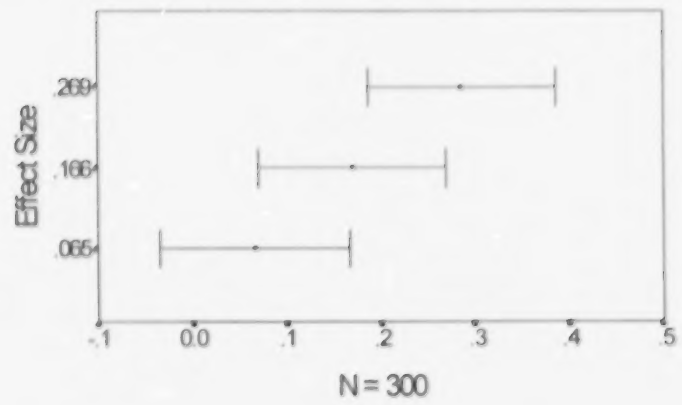
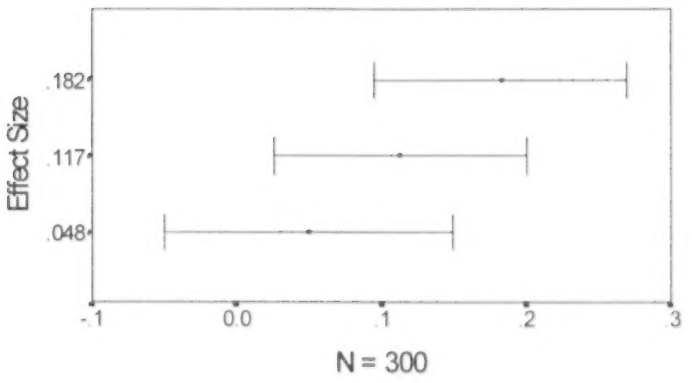


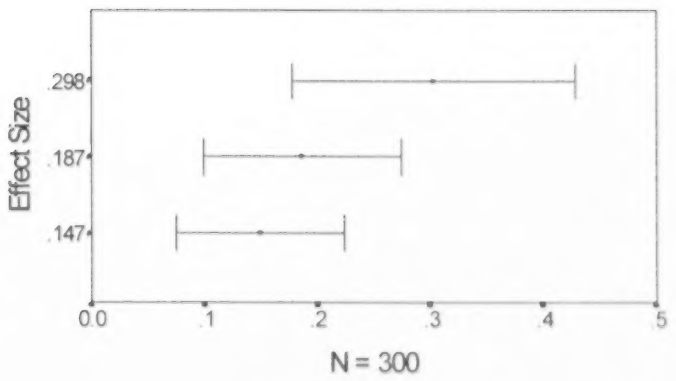
Figure Caption

Figure 4. 3-Point Summary of Linear, Quadratic, and Logistic $\hat{I}$ under Variance Heterogeneity (1:4), Nonnormality (1.75, 3.75), and Large Sample Sizes.

Linear PDA




Quadratic PDA




Logistic